طبقه‌بندی متون علمی به صورت هوشمند

دانشجو: فرهود اطاعتی

استاد راهنما: دکتر بابک نجار اعرابی

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

مقدمه

مقالات علمی به عنوان یکی از مهم‌ترین خروجی‌های پروژه‌های تحقیقاتی و آکادمیک، حاوی اطلاعات متنوع و مهمی می‌باشند و از این جهت، نیاز مبرمی به طبقه‌بندی و دسته‌بندی و در نهایت آنالیز کیفی دارند. از این رو اکثر مجلات و ژورنال‌های معتبر روند داوری کند و وقت‌گیری دارند. ستاد فناوری‌های شناختی ریاست جمهوری، به عنوان یکی از بزرگترین مراکز پژوهش‌های علوم شناختی در ایران، سالانه پذیرای مقالات متعددی در حوزه‌های متفاوت است. با توجه به این که یکی از مراکز اصلی تمرکز ستاد پژوهش در حیطه‌ی زبان‌شناسی است، طبیعی است که در این حیطه مقالات بسیاری را تولید یا جهت چاپ در مجلات خود دریافت نماید، که در هر دو صورت نیازمند آنالیز کیفی و کمی است. از این رو ضروری است که بتوان با سیستمی خودکار به تسهیل این روند پرداخت و اولین قدم در این راه طبقه‌بندی متون دریافتی یا ارسالی است.

هدف از انجام این پروژه ساخت یک سیستم طبقه‌بند برای جداسازی مقالات زبان‌شناسی شناختی از غیرشناختی است به طوری که سیستم بتواند به طور هوشمند و براساس فهم معنای مقالات این دو دسته را از هم جدا نماید. از آن‌جا که قضاوت معنایی مقالات امری خطیر و مهم بوده، سیستم‌های ساخته‌شده برای این امر باید از دقت و ضریب اطمینان بالایی برخوردار باشند تا اعتبار علمی نشریات ستاد خدشه‌دار نگردد.

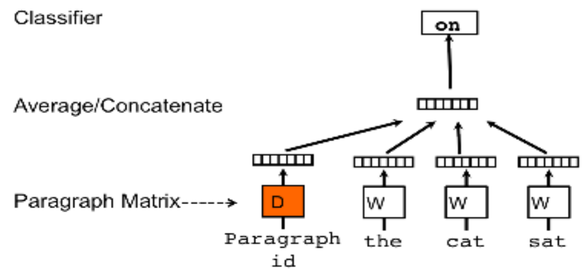
ساختار پیشنهادی

همان‌گونه که در شکل زیر مشخص است برای حل مسئله‌ی طبقه‌بندی لازم است که سه عمل اصلی به‌سازی دادگان دریافتی، استخراج ویژگی از دادگان، و در نهایت طبقه‌بندی دادگان دریافتی انجام گردد.

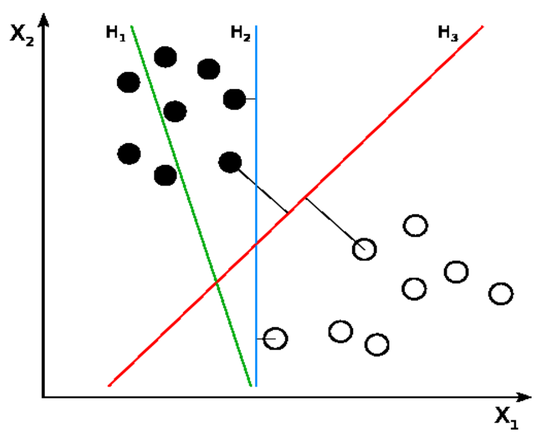
- **به‌سازی متون دریافتی:** در این مرحله متون دریافتی را از کلمات پرتکرار فیلتر می‌کنیم. هم‌چنین جهت کاهش اندازه‌ی لغت‌نامه‌ی ساختار و نویز معنایی کلمات مختلف، مطابق جدول زیر، کلمات هم‌خانواده را به ریشه‌ی مشترک آن‌ها می‌بریم.

Original Word	Lemmatized
Teacher	Teach
Am,is,are	Be

- **ساخت بردار ویژگی:** در این مرحله باید کلمات را از فضای کیفی به فضای کمی و قابل تحلیل ببریم. برای رسیدن به این مهم از نگاشت‌گر متن به بردار Doc2Vec [۲] استفاده نمودیم. در شکل زیر میتوانید شبکه‌ی عصبی مدل D2V را ببینید.



- **طبقه‌بندی بردارهای به دست آمده:** در این مرحله با انتخاب طبقه‌بند ماشین‌بردار پشتیبان[۴] با تابع ریشه‌ی خطی، سعی به پیش‌بینی و طبقه‌بندی بردارهای ویژگی می‌کنیم. در شکل زیر می‌توانید نحوه‌ی کار یک نمونه SVM با تابع ریشه‌ی خطی را ببینید.



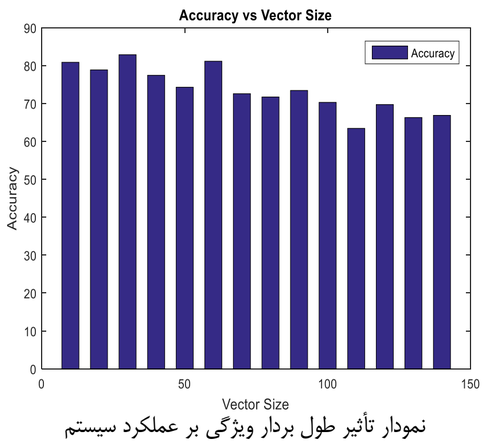
روش ارزیابی نتایج:

در این پروژه با استفاده از روش K-Fold [۱] و با تقسیم کردن فضای داده و تست به نسبت ۱ به ۹ عملکرد سیستم را ارزیابی نمودیم.

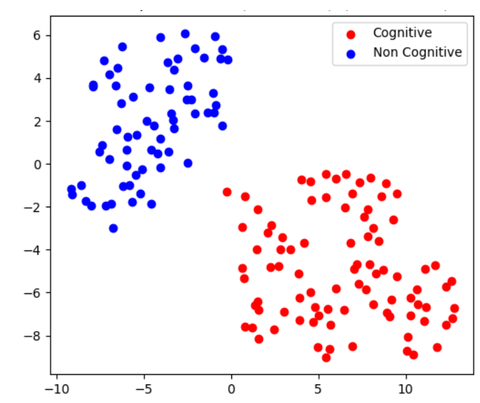
نتایج

همانطور که در اشکال زیر می‌بینید، خصوصیات مدل استخراجگر ویژگی تأثیر مستقیمی بر عملکرد سیستم دارد. با افزایش طول پنجره‌ی D2V عملکرد سیستم افت می‌کند و هم‌چنین با تغییر طول بردار ویژگی نیز می‌توان عملکرد سیستم را تحت تأثیر قرار داد. هم‌چنین در اشکال زیر می‌توانید جداپذیری سیستم طراحی شده را ببینید. در نهایت در جدول زیر می‌توانید عملکرد سیستم را ببینید که نسبت به درصد پروژه‌های قبلی که به ترتیب ٪۷۸ و ٪۸۶ بوده‌اند، پیشرفت داشته است.

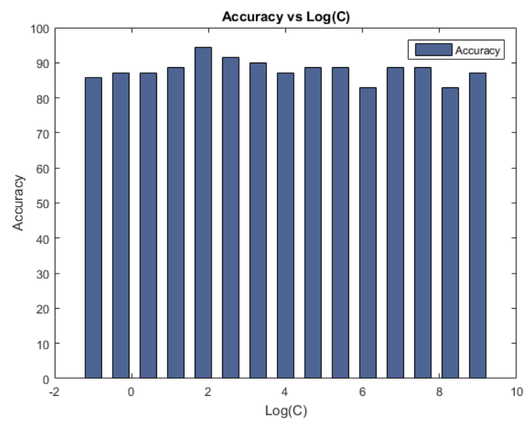
Accuracy	False Positive	False Negative
90.41%	3.68%	5.91%



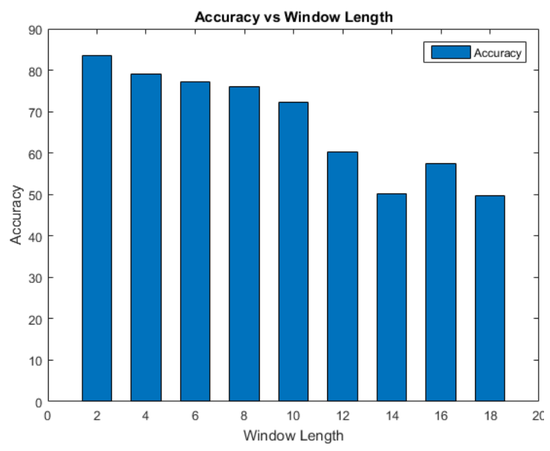
نمودار تأثیر طول بردار ویژگی بر عملکرد سیستم



نمودار جداپذیری بردارهای متن با الگوریتم tSNE [۳]



نمودار تأثیر پارامتر C در دقت سیستم



نمودار تأثیر طول پنجره‌ی D2V بر عملکرد سیستم

جمع بندی

در این پروژه ما موفق به ساخت یک سیستم طبقه‌بند هوشمند متون جهت تشخیص نوع مقالات دریافتی ستاد فناوری‌های علوم‌شناختی ریاست جمهوری شدیم. سیستم طراحی شده پس از پیش‌پردازش متون دریافتی، فضای کلمات را به فضای بردار تبدیل کرده و با استفاده از طبقه‌بند ماشین بردار پشتیبان سعی در طبقه‌بندی متن دریافتی می‌نماید. سیستم پیاده‌سازی شده می‌تواند با دقتی بالای ٪۹۰ نوع مقالات را پیش‌بینی کند که در مقایسه با سیستم‌های طراحی‌شده‌ی قبلی عملکردی کاملاً برتر دارد. در ادامه‌ی مسیر پروژه نیز به سفارش ستاد علوم‌شناختی، یک محیط گرافیکی برای استفاده‌ی آسان کاربران از این سیستم طراحی گشت.

محدودیت‌های پروژه: از محدودیت‌های این پروژه می‌توان به تعداد و تنوع کم داده‌های یادگیری اشاره نمود.

کاربرد های صنعتی و تحقیقاتی: از رویکرد زبان‌شناختی پروژه می‌توان به تحلیل اخبار و شبکه‌های اجتماعی، تحلیل ادبیات مقالات حوزه‌های مختلف جهت داده‌کاوی و استخراج الگوهای پنهان بین توده‌های متون اشاره نمود.

مراجع اصلی

[1] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys,* vol. 4, pp. 40-79, 2010.

[2] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188-1196.

[3] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research,* vol. 9, pp. 2579-2605, 2008.

[4] Wikipedia contributors. (2018). *Support vector machine -- Wikipedia, The Free Encyclopedia*.